

Asymptotic Normality of Random Sums of m -dependent Random Variables

Ümit Işlak

Abstract

We prove a central limit theorem for random sums of the form $\sum_{i=1}^{N_n} X_i$, where $\{X_i\}_{i \geq 1}$ is a stationary m -dependent process and N_n is a random index independent of $\{X_i\}_{i \geq 1}$. Our proof is a generalization of Chen and Shao's result for i.i.d. case and consequently we recover their result. Also a variation of a recent result of Shang on m -dependent sequences is obtained as a corollary. Examples on moving averages and descent processes are provided, and possible applications on non-parametric statistics are discussed.

1 Introduction

In the following, we analyze the asymptotic behavior of random sums of the form $\sum_{i=1}^{N_n} X_i$ as $n \rightarrow \infty$, where X_i 's are non-negative random variables that are stationary and m -dependent, and N_n is a non-negative integer valued random variable independent of X_i 's. Limiting distributions of random sums of independent and identically distributed (i.i.d.) random sums are well studied. See [4], [10], [12] and the references therein. Asymptotic normality of deterministic sums of m -dependent random variables are also well known. See, for example, [2], [9] and [11]. To the best of author's knowledge, previous work on the case of random sums of the form $\sum_{i=1}^{N_n} X_i$ where X_i 's are dependent are limited to [13] where he works on m -dependent random variables and [1] where they investigate random variables that appear as a result of integrating a random field with respect to point processes. Our results here will be in the lines of [4] generalizing their result to the stationary m -dependent case. Throughout the way, we will also improve the results given in [13].

Let's now recall stationary and m -dependent processes. Let $\{X_i\}_{i \geq 1}$ be a stochastic process and let $F_X(X_{i_1+m}, \dots, X_{i_k+m})$ be the cumulative distribution function of the joint distribution of $\{X_i\}_{i \geq 1}$ at times $i_1 + m, \dots, i_k + m$. Then $\{X_i\}_{i \geq 1}$ is said to be *stationary* if, for all k , for all m and for all i_1, \dots, i_k

$$F_X(X_{i_1+m}, \dots, X_{i_k+m}) = F_X(X_{i_1}, \dots, X_{i_k})$$

holds. For more on stationary processes, see [14]. If we define the distance between two subsets of A and B of \mathbb{N} by

$$\rho(A, B) := \inf\{|i - j| : i \in A, j \in B\},$$

then the sequence $\{X_i\}_{i \geq 1}$ is said to be *m -dependent* if $\{X_i, i \in A\}$ and $\{X_j, j \in B\}$ are independent whenever $\rho(A, B) > m$ for $A, B \subset \mathbb{N}$.

An example of a stationary m -dependent process can be given by the moving averages process. Assume that $\{T_i\}_{i \geq 1}$ is a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2 . Letting $X_i = (T_i + T_{i+1})/2$, $\{X_i\}_{i \geq 1}$ is a stationary 1-dependent process with $\mathbb{E}[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2/2$ and $\text{Cov}(X_1, X_2) = \sigma^2/4$.

This paper is organized as follows: In the next section, we state our main results and compare them with previous approaches. In the third section, we give examples on moving averages and descent processes relating it to possible nonparametric tests where the number of observations is itself random. Proofs of the main results are given in Section 4 and we conclude the paper with a discussion of future directions.

2 Main Results

We start with two propositions. Proofs of these are standard and are given at the end of Section 4.

Proposition 2.1. Let $\{X_i\}_{i \geq 1}$ be a stationary m -dependent process with $\mu := \mathbb{E}[X_i]$, $\sigma^2 := \text{Var}(X_i) < \infty$, $a_j := \text{Cov}(X_1, X_{1+j})$. Then for any $N \geq 1$, we have

$$\text{Var} \left(\sum_{i=1}^N X_i \right) = N(\sigma^2 + 2 \sum_{j=1}^m a_j \Gamma_{N,j}) - 2 \sum_{j=1}^m j a_j \Gamma_{N,j}$$

where $\Gamma_{N,j} = \mathbb{1}(N \geq j+1)$.

Proposition 2.2. Let $\{X_i\}_{i \geq 1}$ be as in Proposition 2.1. Let Y_i 's be i.i.d. non-negative integer valued random variables with $\nu := \mathbb{E}[Y_i]$, $\tau^2 := \text{Var}(Y_i) < \infty$ and assume that X_i 's and Y_i 's are independent. Define $N_n = \sum_{i=1}^{N_n} Y_i$. Then we have

$$\text{Var} \left(\sum_{i=1}^{N_n} X_i \right) = n(\nu\sigma^2 + 2\nu \sum_{j=1}^m a_j + \mu^2\tau^2) + \alpha(m)$$

where

$$\alpha(m) = \sum_{k=0}^m (2k \sum_{j=1}^m a_j (\Gamma_{k,j} - 1) - 2 \sum_{j=1}^m j a_j (\Gamma_{k,j} - 1)) - 2 \sum_{j=1}^m j a_j$$

and $\Gamma_{k,j} = \mathbb{1}(k \geq j+1)$. In particular, $\frac{\alpha(m)}{n} \rightarrow 0$ as $n \rightarrow \infty$. When X_i 's are also independent (i.e., $m = 0$), this reduces to

$$\text{Var} \left(\sum_{i=1}^{N_n} X_i \right) = n(\nu\sigma^2 + \mu^2\tau^2).$$

In the following, we will be using \rightarrow_d for convergence in distribution and $=_d$ for equality in distribution. Also $N(0, 1)$ and Φ will denote a standard normal random variable and its cumulative distribution function, respectively. Now we are ready to present our main result.

Theorem 2.3. Let $\{X_i\}_{i \geq 1}$ be a non-negative stationary m -dependent process with $\mu := \mathbb{E}[X_1] > 0$, $\sigma^2 := \text{Var}(X_1) > 0$, $a_j := \text{Cov}(X_1, X_{1+j})$, $\sigma^2 + 2 \sum_{j=1}^m a_j > 0$ and $\mathbb{E}|X_1|^3 < \infty$. Let Y_i 's be i.i.d. non-negative integer valued random variables with $\nu := \mathbb{E}[Y_1] > 0$, $\tau^2 := \text{Var}(Y_1) > 0$, $\mathbb{E}|Y_1|^3 < \infty$ and suppose that X_i 's and Y_i 's are independent. Define $N_n = \sum_{i=1}^{N_n} Y_i$. Then

$$\frac{\sum_{i=1}^{N_n} X_i - n\mu\nu}{\sqrt{n(\nu\sigma^2 + 2\nu \sum_{j=1}^m a_j + \tau^2\mu^2)}} \rightarrow_d N(0, 1) \quad (2.1)$$

as $n \rightarrow \infty$.

Note that assumptions on Y_i 's hold, for example, when Y_i 's are non-degenerate i.i.d. Bernoulli random variables. This is one of the most natural cases as in that case we may consider $\sum_{i=1}^{N_n} X_i$ as the sum of outcomes of a series of experiments, where each observation is blocked with a fixed probability independent of others. The main assumption on X_i 's (others are non-degeneracy conditions) is a third moment condition.

Since our proof is a direct generalization of Chen and Shao's result on i.i.d. case (which is the case with $m = 0$), we recover their result from [4].

Theorem 2.4. Let $\{X_i\}_{i \geq 1}$ be i.i.d. random variables with $\mu := \mathbb{E}[X_1] > 0$, $\sigma^2 := \text{Var}(X_1) > 0$, and assume that $\mathbb{E}|X_1|^3 < \infty$. Let Y_i 's be i.i.d. non-negative integer valued random variables with $\nu := \mathbb{E}[Y_1] > 0$, $\tau^2 := \text{Var}(Y_1) > 0$, $\mathbb{E}|Y_1|^3 < \infty$ and assume that X_i 's and Y_i 's are independent. Define $N_n = \sum_{i=1}^{N_n} Y_i$. Then for any $n \geq 1$, we have

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P} \left(\frac{\sum_{i=1}^{N_n} X_i - n\mu\nu}{\sqrt{n(\nu\sigma^2 + \tau^2\mu^2)}} \leq z \right) - \Phi(z) \right| \leq Cn^{-1/2} \left(\frac{\tau^2}{\nu^2} + \frac{\mathbb{E}[Y_1^3]}{\tau^3} + \frac{\mathbb{E}|X_1|^3}{\nu^{1/2}\sigma^3} + \frac{\sigma}{\mu\sqrt{\nu}} \right) \quad (2.2)$$

where C is a constant independent of n .

We will explain how the proof of Theorem 2.3 also reveals Theorem 2.4 in Section 4. We note that in the original statement of Chen and Shao's result, μ is allowed to be 0. We excluded this in our statement as the upper bound in (2.2) is ∞ when $\mu = 0$.

Our final result will be a variation of the main theorem given in [13] about the asymptotics of random sums of m -dependent random variables. Namely, we have

Theorem 2.5. *Under the assumptions of Theorem 2.3,*

$$\frac{\sum_{i=1}^{N_n} X_i - N_n \mu}{\sqrt{N_n} \left(\sigma + 2 \sum_{j=1}^m a_j \right)} \rightarrow_d N(0, 1) \quad (2.3)$$

as $n \rightarrow \infty$.

Remark 2.6. Indeed, as can be seen from the proof of Theorem 2.3, one can obtain convergence rates when the scaling is perturbed a little bit. More precisely, we have

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P} \left(\frac{\sum_{i=1}^{N_n} X_i - N_n \mu}{\sqrt{N_n} \sigma'} \leq z \right) - \Phi(z) \right| \leq \frac{C}{\sqrt{n}}$$

for a universal constant C and for every $n \geq 1$ where $(\sigma')^2 = \sigma^2 + 2 \sum_{j=1}^m a_j \Gamma_{N_n, j} - \frac{2}{N_n} \sum_{j=1}^m j a_j \Gamma_{N_n, j}$ and $\Gamma_{N_n, j} = \mathbb{1}(N_n \geq j + 1)$.

3 Examples

Example 3.1. (Moving averages) Assume that $\{T_i\}_{i \geq 1}$ is a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2 . Letting

$$X_i = \frac{T_i + T_{i+1}}{2}, \quad i \geq 1,$$

$\{X_i\}_{i \geq 1}$ is a stationary 1-dependent process with $\mathbb{E}[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2/2$ and $\text{Cov}(X_1, X_2) = \sigma^2/4$. When $\mu > 0$ and $\sigma^2 > 0$, we can apply Theorem 2.3 as long as the assumptions on N_n are satisfied (As noted above, they will be satisfied, for example, when Y_i 's are independent Bernoulli random variables with success probability $p \in (0, 1)$). This discussion can be generalized to m -moving averages defined as

$$Y_i = \frac{T_i + T_{i+1} + \dots + T_{i+m-1}}{m}, \quad i \geq 1, \quad m \in \mathbb{N}$$

in a straightforward way.

Example 3.2. (Descent processes) A sequence of real numbers $(t_i)_{i=1}^n$ is said to have a *descent* at position $1 \leq k \leq n-1$ if $t_k > t_{k+1}$. Here we are interested in the descent process of a sequence of random variables. Statistics related to descents are often used in nonparametric statistics to test independence or correlation (For example, one uses the number of inversions in Kendall's tau statistic). See [7] for a brief introduction for this connection. Also see [6] to learn more about why these processes are important.

Now let T_i 's be i.i.d. random variables with distribution F , and $X_i := \mathbb{1}(T_i > T_{i+1})$. Also let Y_i 's be i.i.d. Bernoulli random variables with parameter $p \in (0, 1)$ and set $N_n = \sum_{i=1}^{N_n} Y_i$. Defining

$$W_n = \sum_{j=1}^{N_n-1} X_j,$$

W_n is the number of descents in the random length sequence $(T_1, T_2, \dots, T_{N_n})$.

Here $\{X_i\}_{i \geq 1}$ is a stationary 1-dependent process and it is easy to check that $\mu = 1/2$, $\sigma^2 = 1/4$ and $\sigma^2 + 2a_1 = 1/12$. So assumptions of Theorem 2.3 are satisfied and we obtain the asymptotic normality of W_n .

Example 3.3. (Non-parametric statistics) In this example, we discuss a possible application of Theorem 2.3 in non-parametric statistics. Let T_1, \dots, T_n be the random outcomes of an experiment and assume that the probability of observing any of these is $p \in (0, 1)$ independent of each other. Let N_n be the number of actually observed outcomes and O_1, \dots, O_{N_n} be the corresponding sequence of observations.

Suppose we want to test

$$H_0 : T_1, \dots, T_n \text{ are uncorrelated and } p = p_0.$$

Then one can use the test statistic

$$W_n = \sum_{i=1}^{N_n-1} \mathbb{1}(O_i > O_{i+1})$$

and Theorem 2.3 to understand the asymptotic distribution of W_n under the null hypothesis. A very large or a very small value for this statistics will provide information about the dependence structure of T_i 's.

Extensions of this observation to more general tests will be followed in a subsequent work.

4 Proofs

We start by recalling two results that will be useful in the proof of the main theorem. First of these is a central limit theorem for m -dependent random variables established in [5].

Theorem 4.1. [5] *If $\{X_i\}_{i \geq 1}$ is a sequence of zero mean m -dependent random variables and $W = \sum_{i=1}^n X_i$, then for all $p \in (2, 3]$,*

$$\sup_{z \in \mathbb{R}} |\mathbb{P}(W \leq z) - \Phi(z)| \leq 75(10m + 1)^{p-1} \sum_{i=1}^n \mathbb{E}|X_i|^p.$$

The second result we will need is the following theorem of Chen and Shao ([4]) on the normal approximation of random variables. We note that this theorem is part of what is known as the concentration inequality approach in Stein method literature. See the cited paper or [3] for more on this.

Theorem 4.2. [4] *Let ξ_1, \dots, ξ_n be independent mean zero random variables for $i = 1, \dots, n$ with $\sum_{i=1}^n \text{Var}(\xi_i) = 1$. Let $W = \sum_{i=1}^n \xi_i$, $T = W + \Delta$, and also for each $i = 1, \dots, n$, let Δ_i be a random variable such that ξ_i and $(W - \xi_i, \Delta_i)$ are independent. Then we have*

$$\sup_{z \in \mathbb{R}} |\mathbb{P}(W \leq z) - \Phi(z)| \leq 6.1(\beta_2 + \beta_3) + \mathbb{E}|W\Delta| + \sum_{i=1}^n \mathbb{E}|\xi_i(\Delta - \Delta_i)| \quad (4.1)$$

where

$$\beta_2 = \sum_{i=1}^n \mathbb{E}[\xi_i^2 \mathbb{1}(|\xi_i| > 1)] \quad \text{and} \quad \beta_3 = \sum_{i=1}^n \mathbb{E}[|\xi_i|^3 \mathbb{1}(|\xi_i| \leq 1)].$$

Before moving on to the proof of Theorem 2.3, we finally recall Prokhorov and Kolmogorov distances between probability measures. Let $\mathcal{P}(\mathbb{R})$ be the collection of all probability measures on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ where $\mathfrak{B}(\mathbb{R})$ is the Borel sigma algebra on \mathbb{R} . For a subset $A \subset \mathbb{R}$, define the ϵ -neighborhood of A by

$$A^\epsilon := \{p \in \mathbb{R} : \exists q \in A, d(p, q) < \epsilon\} = \bigcup_{p \in A} B_\epsilon(p)$$

where $B_\epsilon(p)$ is the open ball of radius ϵ centered at p . Then the Prokhorov metric $d_p : \mathcal{P}(\mathbb{R})^2 \rightarrow [0, \infty)$ is defined by setting the distance between two probability measures μ and ν to be

$$d_p(\mu, \nu) := \inf\{\epsilon > 0 : \mu(A) \leq \nu(A^\epsilon) + \epsilon \text{ and } \nu(A) \leq \mu(A^\epsilon) + \epsilon, \forall A \in \mathfrak{B}(\mathbb{R})\}. \quad (4.2)$$

The Kolmogorov distance d_K between two probability measures μ and ν is defined to be

$$d_K(\mu, \nu) = \sup_{z \in \mathbb{R}} |\mu((-\infty, z]) - \nu((-\infty, z])|.$$

The following two facts will be useful: (1) Convergence of measures in Prokhorov metric is equivalent to the weak convergence of measures. (2) Convergence in Kolmogorov distance implies convergence in distribution, but the converse is not true. See, for example, [14] for these standard results.

Now we are ready to prove Theorem 2.3. We will follow the notations of [4] as much as possible.

Proof of Theorem 2.3 : Let Z_1, Z_2 and Z_3 be independent standard normal random variables which are also independent of X_i 's and Y_i 's. Put

$$b = \sqrt{\nu\sigma^2 + 2\nu \sum_{j=1}^m a_j + \tau^2\mu^2}.$$

Define

$$T_n = \frac{\sum_{i=1}^{N_n} X_i - n\mu\nu}{\sqrt{nb}} \quad \text{and} \quad H_n = \frac{\sum_{i=1}^{N_n} X_i - N_n\mu}{\sqrt{N_n}\sigma'}$$

where

$$(\sigma')^2 = \sigma^2 + 2 \sum_{j=1}^m a_j \Gamma_{N_n, j} - \frac{2}{N_n} \sum_{j=1}^m j a_j \Gamma_{N_n, j} \quad (4.3)$$

with $\Gamma_{N_n, j} := \mathbb{1}(N_n \geq j + 1)$. Also write

$$T_n = \frac{\sqrt{N_n}\sigma'}{\sqrt{nb}} H_n + \frac{(N_n - n\nu)\mu}{\sqrt{nb}}$$

and

$$T_n(Z_1) = \frac{\sqrt{N_n}\sigma'}{\sqrt{nb}} Z_1 + \frac{(N_n - n\nu)\mu}{\sqrt{nb}}.$$

For n large enough, we have $m < n\nu/2$. For such n , we have

$$\begin{aligned} d_K(T_n, T_n(Z_1)) &= d_K(H_n, Z_1) \\ &\leq \mathbb{P}(|N_n - n\nu| > n\nu/2) + \sup_{z \in \mathbb{R}} \mathbb{E}[\mathbb{E}[|\mathbb{1}(H_n \leq z) - \mathbb{1}(Z_1 \leq z)| \mathbb{1}(|N_n - n\nu| \leq n\nu/2) | N_n]] \\ &\leq \frac{4\tau^2}{n\nu^2} + \mathbb{E} \left[\mathbb{E} \left[\frac{CN_n \mathbb{E}|X_1|^3 \mathbb{1}(|N_n - n\nu| \leq n\nu/2)}{N_n^{3/2} \left(\sigma^2 + 2 \sum_{j=1}^m a_j - \frac{2}{N_n} \sum_{j=1}^m j a_j \right)^{3/2}} | N_n \right] \right] \end{aligned} \quad (4.4)$$

where for (4.4) we used Chebyshev's inequality for the first estimate and Theorem 4.1 with $p = 3$ for the second estimate. Here the condition that $m < n\nu/2$ simplifies $(\sigma')^2$ as defined in (4.3) to $(\sigma')^2 = \left(\sigma^2 + 2 \sum_{j=1}^m a_j - \frac{2}{N_n} \sum_{j=1}^m j a_j \right)$ when $|N_n - n\nu| \leq n\nu/2$. Also note that throughout this proof, C will be a positive constant with not necessarily the same value in different lines. Now if $\sum_{j=1}^m j a_j < 0$, then the bound in (4.4) yields

$$d_K(T_n, T_n(Z_1)) \leq \frac{4\tau^2}{n\nu^2} + \frac{C\mathbb{E}|X_1|^3}{\sqrt{n\nu/2} \left(\sigma^2 + 2 \sum_{j=1}^m a_j \right)^{3/2}} \rightarrow 0$$

as $n \rightarrow \infty$. Else if $\sum_{j=1}^m j a_j \geq 0$, we observe that for large enough n , we have $\sigma^2 + 2 \sum_{j=1}^m a_j - \frac{4}{n\nu} \sum_{j=1}^m j a_j > 0$ by our assumption that $\sigma^2 + 2 \sum_{j=1}^m a_j > 0$. For such n , using the bound in (4.4) we obtain

$$d_K(T_n, T_n(Z_1)) \leq \frac{4\tau^2}{n\nu^2} + \frac{C\mathbb{E}|X_1|^3}{\sqrt{n\nu/2} \left(\sigma^2 + 2 \sum_{j=1}^m a_j - \frac{4}{n\nu} \sum_{j=1}^m j a_j \right)^{3/2}} \quad (4.5)$$

and this yields $d_K(T_n, T_n(Z_1)) \rightarrow 0$ as $n \rightarrow \infty$ when $\sum_{j=1}^m j a_j \geq 0$.

Hence we conclude that $d_K(T_n, T_n(Z_1)) \rightarrow 0$ as $n \rightarrow \infty$ as long as $\nu > 0$ and $\sigma^2 + 2 \sum_{j=1}^m a_j > 0$. This in particular implies

$$d_p(T_n, T_n(Z_1)) \rightarrow 0 \quad (4.6)$$

as $n \rightarrow \infty$ where d_P is the Prokhorov distance as defined in (4.2).

Next let $(\sigma'')^2 = (\sigma')^2 + 2 \sum_{j=1}^m a_j(1 - \Gamma_{N_n, j}) + \frac{2}{N_n} \sum_{j=1}^m j a_j \Gamma_{N_n, j}$ so that

$$(\sigma'')^2 = \sigma^2 + 2 \sum_{j=1}^m a_j.$$

Note that σ'' is not random and introduce

$$T'_n(Z_1) = \frac{\sqrt{N_n} \sigma''}{\sqrt{nb}} Z_1 + \frac{(N_n - n\nu)\mu}{\sqrt{nb}}$$

and

$$T_n(Z_1, Z_2) := \frac{\tau\mu}{b} \left(Z_2 + \frac{\sigma'' \sqrt{\nu}}{\tau\mu} Z_1 \right).$$

One can easily check that $T_n(Z_1, Z_2)$ is a standard normal random variable since Z_1 and Z_2 are assumed to be independent. So if we can show that $d_p(T_n(Z_1), T'_n(Z_1)) \rightarrow 0$ and $d_p(T'_n(Z_1), T_n(Z_1, Z_2)) \rightarrow 0$ as $n \rightarrow \infty$, then the result will follow from an application of triangle inequality. We start by showing that $d_p(T'_n(Z_1), T_n(Z_1, Z_2)) \rightarrow 0$. For this purpose, we will use Chen-Shao's concentration inequality approach to get bounds in the Kolmogorov distance and to recover Chen and Shao's result on i.i.d. case (If we just wanted to show $d_P(T'_n(Z_1), T_n(Z_1, Z_2)) \rightarrow 0$, then this could be done in a much easier way. See Remark 4.3). The following argument is in a sense rewriting the corresponding proof in [4] with slight changes since the concentration approach is used on N_n which is in both problems a sum of independent random variables. For the sake of completeness, we include all details.

Define the truncation \bar{x} of $x \in \mathbb{R}$ by

$$\bar{x} = \begin{cases} n\nu/2 & \text{if } x < n\nu/2 \\ x & \text{if } n\nu/2 \leq x \leq 3n\nu/2 \\ 3n\nu/2 & \text{if } x > 3n\nu/2 \end{cases}$$

and let

$$\bar{T}'_n = \frac{\sqrt{N_n} \sigma''}{\sqrt{nb}} Z_1 + \frac{(N_n - n\nu)\mu}{\sqrt{nb}} = \frac{\tau\mu}{b} \left(W + \Delta + \frac{\sigma'' \sqrt{\nu}}{\tau\mu} Z_1 \right)$$

where

$$W = \frac{N_n - n\nu}{\sqrt{n}\tau} \quad \text{and} \quad \Delta = \frac{(\sqrt{N_n} - \sqrt{n\nu})\sigma'' Z_1}{\sqrt{n}\tau\mu}.$$

Since Y_i is independent of $N_n - Y_i$ for all $i = 1, \dots, n$, we can apply Theorem 4.2 to $W + \Delta$ setting

$$\Delta_i = \frac{\sqrt{N_n - Y_i + \nu} - \sqrt{n\nu} \sigma'' Z_1}{\sqrt{n}\tau\mu}, \quad i = 1, \dots, n.$$

(So $\xi_i = \frac{Y_i - \nu}{\sqrt{n}\tau}$ in Theorem 4.2.) For the first term of the upper bound given in (4.1), we have

$$6.1(\beta_2 + \beta_3) \leq 6.1(2n) \mathbb{E} \left| \frac{Y_1}{\sqrt{n}\tau} \right|^3 \leq \frac{Cn \mathbb{E}|Y_1|^3}{(n\tau^2)^{3/2}} = \frac{C \mathbb{E}|Y_1|^3}{\tau^3 \sqrt{n}}. \quad (4.7)$$

For the second term in (4.1), we have

$$\mathbb{E}|W\Delta| = \mathbb{E}|Z_1|\mathbb{E}\left[\frac{\sigma''}{\sqrt{n\tau\mu}}\mathbb{E}|W(\sqrt{\overline{N_n}} - \sqrt{n\nu})|\right] = \frac{\mathbb{E}|Z_1|\sigma''}{\sqrt{n\tau\mu}}\mathbb{E}\left|W\frac{\overline{N_n} - n\nu}{\sqrt{\overline{N_n}} + \sqrt{n\nu}}\right|$$

where we used the identity $\sqrt{x} - \sqrt{y} = \frac{x-y}{\sqrt{x}+\sqrt{y}}$ in the second equality. So by an application of Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \mathbb{E}|W\Delta| &\leq \frac{C\sigma''}{\sqrt{n\tau\mu}}(\mathbb{E}|W|^2)^{1/2} \left(\mathbb{E}\left|\frac{\overline{N_n} - n\nu}{\sqrt{\overline{N_n}} + \sqrt{n\nu}}\right|^2 \right)^{1/2} \leq \frac{C\sigma''}{\sqrt{n\tau\mu}} \left(\mathbb{E}\left|\frac{N_n - n\nu}{\sqrt{n\nu}}\right|^2 \right)^{1/2} \\ &\leq \frac{C\sigma''}{\sqrt{n\nu\mu}}. \end{aligned} \quad (4.8)$$

since $\mathbb{E}[W^2] = 1$ and $\text{Var}(N_n) = n\tau^2$. Also note that for the second inequality we used $|\overline{N_n} - n\nu| \leq |N_n - n\nu|$ which easily from the definition of the truncation.

For the third term of the bound in (4.1), we have

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}|\xi_i(\Delta - \Delta_i)| &\leq \sum_{i=1}^n (\mathbb{E}|\xi_i|^2)^{1/2} \mathbb{E}(|\Delta - \Delta_i|^2)^{1/2} \leq \sum_{i=1}^n \frac{1}{\sqrt{n}} \left(\mathbb{E}\left|\left(\frac{\sqrt{\overline{N_n}} - \sqrt{\overline{N_n} - Y_i + \nu}}{\sqrt{n\tau\mu}}\right)\sigma''Z_1\right|^2 \right)^{1/2} \\ &\leq n \frac{\mathbb{E}|Z_1|\sigma''}{\sqrt{n}} \left(\mathbb{E}\left|\frac{\sqrt{\overline{N_n}} - \sqrt{\overline{N_n} - Y_1 + \nu}}{\sqrt{n\tau\mu}}\right|^2 \right)^{1/2} \\ &\leq C\sqrt{n}\sigma'' \left(\mathbb{E}\left|\frac{\overline{N_n} - \overline{N_n} - Y_1 + \nu}{\sqrt{n\tau\mu}(\sqrt{\overline{N_n}} + \sqrt{\overline{N_n} - Y_1 + \nu})}\right|^2 \right)^{1/2} \\ &\leq \frac{C\sigma''}{\tau\mu} \frac{(\mathbb{E}|Y_1 - \nu|^2)^{1/2}}{\sqrt{n\nu/2} + \sqrt{n\nu/2}} \end{aligned}$$

where we used $\mathbb{E}|\xi_i|^2 = 1/n$, the identity $\sqrt{x} - \sqrt{y} = \frac{x-y}{\sqrt{x}+\sqrt{y}}$ and the inequality $|\overline{x} - \overline{x - y}| \leq |y|$.

We conclude

$$\sum_{i=1}^n \mathbb{E}|\xi_i(\Delta - \Delta_i)| = \sum_{i=1}^n \mathbb{E}\left|\frac{Y_i - \nu}{\sqrt{n\tau^2}}(\Delta - \Delta_i)\right| \leq \frac{C\sigma''}{\sqrt{n\nu\mu}}. \quad (4.9)$$

Using Theorem 4.2, we get

$$\begin{aligned} \sup_{z \in \mathbb{R}} |\mathbb{P}(T'_n(Z_1) \leq z) - \mathbb{P}(T_n(Z_1, Z_2) \leq z)| &\leq \mathbb{P}(|N_n - n\nu| > n\nu/2) \\ &+ \sup_{z \in \mathbb{R}} \mathbb{E}[\mathbb{E}[\mathbb{1}(\overline{T}'_n(Z_1) \leq z) - \mathbb{1}(T_n(Z_1, Z_2) \leq z)] \mathbb{1}(|N_n - n\nu| \leq n\nu/2) | N_n] \\ &\leq \sup_{z \in \mathbb{R}} |\mathbb{P}(W + \Delta \leq z) - \mathbb{P}(Z_3 \leq z)| \\ &\leq \frac{4\tau^2}{n\nu^2} + C \left(\frac{|Y_1|^3}{\tau^3\sqrt{n}} + \frac{\sigma''}{\sqrt{n\nu\mu}} \right). \end{aligned} \quad (4.10)$$

where for the last step we combined the three estimates given in (4.7), (4.8) and (4.9). Thus,

$$d_p(T'_n(Z_1), T_n(Z_1, Z_2)) \longrightarrow 0 \quad (4.11)$$

as $n \rightarrow \infty$ if $\nu, \tau, \mu > 0$.

Finally we need to show that $d_P(T_n(Z_1), T'_n(Z_1)) \rightarrow 0$. First observe that

$$T_n(Z_1) - T'_n(Z_1) = \frac{\sqrt{N_n}(\sigma' - \sigma'')Z_1}{\sqrt{nb}} \rightarrow 0$$

almost surely as $n \rightarrow \infty$. Also we know that $T'_n(Z_1)$ converges in distribution to $T_n(Z_1, Z_2)$. Thus, using Slutsky's theorem we conclude that $T_n(Z_1) = T'_n(Z_1) + T_n(Z_1) - T'_n(Z_1)$ also converges in distribution to $T_n(Z_1, Z_2)$. Hence

$$d_P(T_n(Z_1), T'_n(Z_1)) \leq d_P(T_n(Z_1), T_n(Z_1, Z_2)) + d_P(T_n(Z_1, Z_2), T'_n(Z_1)) \rightarrow 0 \quad (4.12)$$

as $n \rightarrow \infty$.

Hence combining (4.6), (4.11) and (4.12), we obtain

$$d_p(T_n, T_n(Z_1, Z_2)) \leq d_p(T_n, T_n(Z_1)) + d_p(T_n(Z_1), T'_n(Z_1)) + d_p(T'_n(Z_1), T_n(Z_1, Z_2)) \rightarrow 0$$

as $n \rightarrow \infty$ under the given assumptions and result follows. \square

Remark 4.3. We can show that $d_P(T'_n(Z_1), T_n(Z_1, Z_2)) \rightarrow 0$ easily if we are not interested in convergence rates. To see this, note that we can write $T'_n(Z_1)$ as

$$T'_n(Z_1) = \sqrt{\frac{N_n}{n}} \frac{\sigma''}{b} \left(Z_1 + \frac{\frac{(N_n - n\nu)}{\sqrt{n\tau}} \frac{\mu\tau}{b}}{\sqrt{\frac{N_n}{n}} \frac{\sigma''}{b}} \right).$$

Now by the strong law of large numbers $\frac{N_n}{n} \rightarrow \nu$ a.s. and by the standard central limit theorem for independent random variables $\frac{N_n - n\nu}{\sqrt{n\tau}} \rightarrow Z$ where Z is a standard normal random variable independent of Z_1 . Using Slutsky's theorem twice with these observations immediately reveals that $T'_n(Z_1)$ converges in distribution to a standard normal random variable.

Proof of Theorem 2.4 : First note that under independence, we have $a_j = 0$ for $j = 1, \dots, m$ so that $\sigma' = \sigma'' = \sigma$. Following the proof of Theorem 2.3, this implies that $d_K(T_n(Z_1), T'_n(Z_1)) = 0$ for every n . Now the result follows from the estimates of $d_K(T_n, T_n(Z_1))$ and $d_K(T_n(Z_1), T_n(Z_1, Z_2))$ by substituting $a_j = 0$ for $j = 1, \dots, m$. \square

Proof of Corollary 2.5 : In the proof of Theorem 2.3, we showed that

$$d_K(H_n, Z_1) = d_K(T_n, T_n(Z_1)) \rightarrow 0$$

where $H_n = \frac{\sum_{i=1}^{N_n} X_i - N_n\mu}{\sqrt{N_n}\sigma'}$ and $(\sigma')^2 = \sigma^2 + 2 \sum_{j=1}^m a_j \Gamma_{N_n, j} - \frac{2}{N_n} \sum_{j=1}^m j a_j$. Since $\frac{\sigma'}{\sqrt{\sigma^2 + 2 \sum_{j=1}^m a_j}} \rightarrow 1$ a.s., result follows from Slutsky's theorem. \square

Finally we give the proofs of the variance formulas given in Proposition 2.1 and 2.2.

Proof of Proposition 2.1 : We have

$$\begin{aligned} \text{Var} \left(\sum_{i=1}^N X_i \right) &= \sum_{i=1}^N \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq N} \text{Cov}(X_i, X_j) \\ &= N\sigma^2 + 2 \sum_{j=1}^m (N-j) a_j \mathbb{1}(N \geq j+1) \end{aligned}$$

Rearranging terms, we obtain

$$\text{Var} \left(\sum_{i=1}^N X_i \right) = N \left(\sigma^2 + 2 \sum_{j=1}^m a_j \mathbb{1}(N \geq j+1) \right) - 2 \sum_{j=1}^m j a_j \mathbb{1}(N \geq j+1)$$

by which the variance formula follows. \square

Proof of Proposition 2.2 : First note that assumptions of Wald's identity are satisfied and so $\mathbb{E} \left[\sum_{i=1}^{N_n} X_i \right] = n\nu\mu$. Using this, we get

$$\begin{aligned} \text{Var} \left(\sum_{i=1}^{N_n} X_i \right) &= \mathbb{E} \left[\left(\sum_{i=1}^{N_n} X_i - n\nu\mu \right)^2 \right] \\ &= \sum_{k=m+1}^{\infty} \mathbb{E} \left(\sum_{i=1}^k X_i - n\nu\mu \right)^2 \mathbb{P}(N_n = k) + \sum_{k=0}^m \mathbb{E} \left(\sum_{i=1}^k X_i - n\nu\mu \right)^2 \mathbb{P}(N_n = k) \end{aligned}$$

where for the second equality we conditioned on N_n which is independent of X_i 's. Next note that we have

$$\mathbb{E} \left[\sum_{i=1}^k X_i \right] = k\mu \quad \text{and} \quad \mathbb{E} \left(\sum_{i=1}^k X_i \right)^2 = k\sigma^2 + 2k \sum_{j=1}^m a_j \Gamma_{j,k} - 2 \sum_{j=1}^m j a_j \Gamma_{j,k} + k^2 \mu^2 \quad (4.13)$$

with $\Gamma_{j,k} = \mathbb{1}(j \geq k+1)$. Thus, using Proposition 2.1 and (4.13), and doing some elementary manipulations, we obtain

$$\begin{aligned} \text{Var} \left(\sum_{i=1}^{N_n} X_i \right) &= \sum_{k=m+1}^{\infty} \mathbb{E} \left(\sum_{i=1}^k X_i - k\mu + k\mu - n\nu\mu \right)^2 \mathbb{P}(N_n = k) \\ &+ \sum_{k=0}^m \mathbb{E} \left[\left(\sum_{i=1}^k X_i \right)^2 - 2n\nu\mu \left(\mathbb{E} \left[\sum_{i=1}^k X_i \right] \right) + n^2 \nu^2 \mu^2 \right] \mathbb{P}(N_n = k) \\ &= \sum_{k=m+1}^{\infty} \left(\text{Var} \left(\sum_{i=1}^k X_i \right) + (k\mu - n\nu\mu)^2 \right) \mathbb{P}(N_n = k) \\ &+ \sum_{k=0}^m (k\sigma^2 + 2k \sum_{j=1}^m a_j \Gamma_{j,k} - 2 \sum_{j=1}^m j a_j \Gamma_{j,k} + k^2 \mu^2 - 2n\nu\mu^2 k + n^2 \nu^2 \mu^2) \mathbb{P}(N_n = k) \end{aligned}$$

Noting that for $k \geq m+1$, $\text{Var} \left(\sum_{i=1}^k X_i \right) = k \left(\sigma^2 + 2 \sum_{j=1}^m a_j \right) - 2 \sum_{j=1}^m j a_j$, we get

$$\begin{aligned} \text{Var} \left(\sum_{i=1}^{N_n} X_i \right) &= \sum_{k=0}^{\infty} (k\sigma^2 + 2k \sum_{j=1}^m a_j - 2 \sum_{j=1}^m j a_j + k^2 \mu^2 - 2kn\nu\mu^2 + n^2 \nu^2 \mu^2) \mathbb{P}(N_n = k) \\ &+ \sum_{k=0}^m (k\sigma^2 + 2k \sum_{j=1}^m a_j \Gamma_{j,k} - 2 \sum_{j=1}^m j a_j \Gamma_{j,k} + \mu^2 k^2 - 2n\nu\mu^2 k + n^2 \nu^2 \mu^2 \\ &- k\sigma^2 - 2k \sum_{j=1}^m a_j + 2 \sum_{j=1}^m j a_j - k^2 \mu^2 + 2kn\nu\mu^2 - n^2 \nu^2 \mu^2) \mathbb{P}(N_n = k). \end{aligned}$$

After some cancelations and using the values for $\mathbb{E}[N_n] = n\nu$ and $\mathbb{E}[N_n^2] = n\tau^2 + n^2\nu^2$, we finally arrive at

$$\text{Var} \left(\sum_{i=1}^{N_n} X_i \right) = n(\nu\sigma^2 + 2\nu \sum_{j=1}^m a_j + \mu^2 \tau^2) + \alpha(m)$$

where

$$\alpha(m) = \sum_{k=0}^m (2k \sum_{j=1}^m a_j (\Gamma_{k,j} - 1) - 2 \sum_{j=1}^m j a_j (\Gamma_{k,j} - 1)) - 2 \sum_{j=1}^m j a_j.$$

The assertion that $\frac{\alpha(m)}{n} \rightarrow 0$ as $n \rightarrow \infty$ follows from the fact that all the variables are bounded. \square

5 Conclusion

In this paper, we established a central limit theorem for random sums of stationary m -dependent processes. Our proof is an extension of the argument given in [4] for the i.i.d. case and this enables to recover their result. At the same time, we were able to give variations of the results in [13]. In the subsequent research we are planning to (1) obtain convergence rates for Theorem 2.3, (2) relax the m -dependence condition to a weak local dependence condition (For such conditions, see [5]), (3) adapt the size biasing technique often used in normal approximation to the case of random sums (See, for example, [8]) and (4) find more applications on non-parametric statistics.

References

- [1] Barbour, A. D. and Xia, A., (2006). *Normal approximation for random sums*, Adv. in Appl. Probab. 38, no. 3, 693-728.
- [2] Bergström, H., (1970). *A comparison method for distribution functions of sums of independent and dependent random variables*. Teor. Verojatnost. i Primenen. 15 442-468.
- [3] Chen L.H.Y. , Goldstein L. and Shao Q. M., (2011). *Normal approximation by Stein's method*. Springer; Berlin, Heidelberg.
- [4] Chen, L. and Shao, Q., (2007). *Normal approximation for nonlinear statistics using a concentration inequality approach*, Bernoulli 13 581-599.
- [5] Chen, L. H. Y. and Shao, Q. M., (2004). *Normal approximation under local dependence*, Ann. Prob. 32, 1985-2028.
- [6] P. Diaconis, A. Borodin and J. Fulman, (2009). *On adding a list of numbers (and other one-dependent determinantal processes)*. Bulletin (New Series) of the Amer. Math. Soc., 47(4):639-670.
- [7] Ferguson, Thomas S., (1996). *A course in large sample theory*, Texts in Statistical Science Series, Chapman & Hall, London.
- [8] Goldstein, L., (2005). *Berry Esseen Bounds for Combinatorial Central Limit Theorems and Pattern Occurrences, using Zero and Size Biasing*, Journal of Applied Probability, vol 42, pp. 661-683.
- [9] Hoeffding, W. and Robbins, H., (1948). *The central limit theorem for dependent random variables*. Duke Math. J. 15, 773-780.
- [10] Kläver, H. and Schmitz, N., (2006). *An inequality for the asymmetry of distributions and a Berry-Esseen theorem for random summation*, J. Inequal. Pure Appl. Math. 7, no. 1, Article 2, 12 pp.
- [11] Orey, S., (1958). *A central limit theorem for m -dependent random variables*. Duke Math. J., 25, 543-546.
- [12] Robbins, H., (1948). *The asymptotic distribution of the sum of a random number of random variables*. Bull. Amer. Math. Soc. 54, 1151-1161.
- [13] Shang, Y., (2012). *A central limit theorem for randomly indexed m -dependent random variables*, Filomat 26:4, 713-717.
- [14] Shiryaev, A. N., (1996). *Probability*, Second edition. Graduate Texts in Mathematics, 95. Springer-Verlag, New York, xvi+623 pp.